33 Bits of Entropy: Myths and Fallacies of "Personally Identifiable Information"

Arvind Narayanan
Princeton University

ABSTRACT

Data is the currency of the digital economy, but increasing data collection by companies and sharing with third parties threatens privacy. "Anonymization" is the usual answer to privacy concerns, typically implemented via removal of "personally identifiable information."

Sweeney's work on reidentification of Massachusetts hospital records showed that naive deidentification via PII removal can be reversed [3]. That led to a cat-and-mouse game between deidentification and reidentification, with standards such as HIPAA mandating removal of a more comprehensive set of attributes. In parallel, techniques for data transformations that enable *specific* categories of computations in a mathematically rigorous privacy-preserving way were developed — "differential privacy" enables sidestepping the need for anonymization altogether [1].

However, the anonymization paradigm is extremely popular due to its convenience and because it avoids the need to circumscribe allowed computations in advance. Several theoretical and practical questions remained open. Given the increasingly easy availability of public "auxiliary information" about individuals (e.g., from social media), is it possible to provide any technical privacy guarantees via anonymization, while maintaining data utility? How identifiable are people's footprints in the rich "longitudinal" databases that are common today? Can we characterize which types of data can lead to reidentification, thus salvaging the notion of "Personally Identifiable Information?" Finally, how many bits of uncorrelated information ("entropy") are required to reidentify individuals in large datasets?

This paper references the recent work, *Myths and Fallacies of "Personally Identifiable Information"* [2].

BODY

Anonymization of rich consumer data is infeasible—people are unique, and any piece of data can help reidentify. 33 bits of entropy will do.

REFERENCES

[1] C. Dwork. Differential privacy: A survey of results. In *TAMC*, pages 1–19, 2008.

[2] A. Narayanan and V. Shmatikov. Myths and fallacies of "personally identifiable information". *Commun. ACM*, 53(6):24–26, June 2010.

[3] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

Volume 1 of Tiny Transactions on Computer Science